



# International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





# Fake Reviews Detection using Supervised Machine Learning Techniques

S. Venkatesan<sup>1</sup>, K. Madhavi<sup>2</sup>, V. Greeshma<sup>3</sup>, B. Ruthika<sup>4</sup>, K. Divya<sup>5</sup>, K. Madhusudhan<sup>6</sup>,  
V. Karthik<sup>7</sup>

Associate Professor, Department of CSE (Data Science), NSRIT, Vishakhapatnam, India<sup>1,2</sup>

Student of Department of CSE (Data Science), NSRIT, Vishakhapatnam, India<sup>3,4,5,6,7</sup>

**ABSTRACT:** Online review systems have become an integral component of modern digital marketplaces, shaping consumer perceptions and strongly influencing purchasing decisions. Platforms such as Yelp and Amazon rely heavily on user-generated content to establish trust, credibility, and transparency between consumers and businesses. However, the growing economic and reputational impact of online reviews has also encouraged malicious behaviours, particularly the generation of deceptive or fake reviews intended to manipulate public opinion. Detecting such reviews is a challenging yet essential task within the fields of natural language processing (NLP) and machine learning.

This study presents an expanded and comprehensive supervised machine learning framework for detecting fake reviews using textual data and limited metadata. The proposed approach evaluates multiple classical machine learning algorithms—including Naive Bayes, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP)—on a labelled Yelp-like dataset. Reviews are pre-processed using standard NLP techniques and represented using a bag-of-words model via Count Vectorizer.

**KEYWORDS:** Online reviews, Fake reviews, Supervised Machine Learning, Language patterns, Logistic regression, Support Vector Machines (SVM).

## I. INTRODUCTION

The exponential growth of e-commerce and online service platforms has fundamentally transformed how consumers evaluate products and services. Rather than relying solely on advertisements or personal recommendations, users increasingly depend on online reviews to inform their decisions. Studies have shown that even small changes in average ratings can significantly impact consumer behaviour and business revenue. As a result, review platforms have become high-value targets for manipulation.

Fake reviews—also referred to as deceptive opinion spam—are deliberately crafted to mislead readers by presenting biased or false experiences. These reviews may be generated by competitors seeking to damage reputations, businesses attempting to inflate their own ratings, or paid individuals and automated bots. The detection of such reviews is difficult due to their linguistic similarity to genuine content and the evolving strategies used by spammers.

Fake review detection is commonly modelled as a supervised text classification problem, where labelled examples of genuine and fake reviews are used to train predictive models. Prior research has explored a wide range of approaches, including linguistic feature engineering, sentiment analysis, behavioural modelling, and deep learning. Despite recent advances in neural networks and transformer-based models, classical supervised machine learning methods remain attractive due to their interpretability, efficiency, and strong performance on structured textual data.



## II. METHODOLOGY

The design and implementation of an AI-based fake review detection system require a balanced integration of technical robustness and ethical responsibility. Since online reviews play a critical role in shaping consumer opinions, influencing purchasing behavior, and affecting business reputation, any automated system that analyses or classifies such content must be designed with accuracy, fairness, transparency, and accountability in mind. The proposed system employs supervised machine learning techniques to classify reviews as genuine or fake while ensuring responsible AI usage, ethical data handling, and secure system operation.

### 2.1. Data Privacy and Confidentiality

#### 2.1.1. Ethical Use of User-Generated Review Data

Online reviews are voluntarily written by users and often express personal experiences, emotions, and opinions. Although the dataset used in this project consists of publicly available reviews, ethical considerations require that such data be handled responsibly. The system processes review content strictly for analytical and research purposes and does not attempt to infer or expose personal user identities. The objective is to detect deceptive linguistic patterns rather than evaluate or judge individual reviewers.

#### 2.2.2. Data Minimization and Anonymization

To further enhance privacy protection, the system follows the principle of data minimization. Only essential attributes, such as review text and associated ratings, are used for model training and evaluation. Any metadata related to user identity, geographical location, or account history is excluded from the learning process. This ensures that predictions are derived solely from textual characteristics, reducing the risk of profiling or discriminatory outcomes.

#### 2.2.3. Secure Storage and Responsible Data Handling

Datasets and trained models are stored in secure environments with restricted access to prevent unauthorized usage. Intermediate files created during preprocessing, such as cleaned text or vectorized features, are treated as temporary artifacts and removed after processing. These practices minimize data retention risks and align with ethical data governance standards.

### 2.2. Accountability in AI Decision-Making

#### 2.2.1. Responsibility in Automated Review Classification

While supervised machine learning models can effectively identify deceptive patterns, they are not infallible. The system may occasionally misclassify reviews due to ambiguity, linguistic complexity, or insufficient contextual information. Therefore, the fake review detection system is designed as a decision-support mechanism rather than an autonomous decision-maker. Human moderators or analysts are encouraged to review flagged reviews before enforcing moderation actions.

#### 2.2.2. Transparency of Model Behavior

Machine learning models, particularly ensemble and neural network models, may exhibit limited interpretability. To address this challenge, the system includes interpretable classifiers such as Multinomial Naive Bayes and Decision Trees alongside more complex models. Evaluation metrics, confusion matrices, and comparative performance analyses are documented to provide insights into model behavior and limitations.

#### 2.2.3. Governance and Ethical Usage Guidelines

Organizations deploying fake review detection systems should establish clear governance frameworks that define acceptable usage, limitations of automation, and procedures for handling misclassifications. Ethical guidelines help prevent misuse of predictions, such as unfair penalization of users or businesses, and ensure that automated tools are applied responsibly.

### 2.3 Accuracy, Bias, and Fairness in AI Models

#### 2.3.1 Linguistic Challenges in Fake Review Detection

Fake reviews often mimic genuine language patterns, making detection a complex task. Informal writing styles, sarcasm, grammatical errors, and short review lengths can reduce classification accuracy. Additionally, cultural and contextual differences in language usage may affect model performance. To address these challenges, the system



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

employs comprehensive text preprocessing and continuous performance evaluation.

### 2.3.2 Bias in Training Data and Model Learning

Supervised learning models inherit biases present in the training dataset. For example, an imbalance between positive and negative reviews may lead to skewed predictions. To mitigate such bias, the dataset is analysed for class distribution, and performance is evaluated using precision, recall, and F1-score rather than relying solely on accuracy.

### 2.3.3 Fair Treatment of Genuine Reviews

False positives—where genuine reviews are incorrectly classified as fake—can negatively impact both users and businesses. Therefore, the system prioritizes balanced classification performance. Special attention is given to minimizing false positives while maintaining effective detection of deceptive reviews.

## III. SYSTEM EFFICIENCY AND PERFORMANCE

### 3.1 Computational Resource Optimization

The system is designed to operate efficiently on commonly available hardware. Feature extraction using bag-of-words representation combined with classical supervised machine learning models ensures low computational overhead while delivering reliable performance. This design choice enables practical deployment without requiring specialized hardware.

### 3.2 Scalability and Processing Efficiency

Efficient vectorization and optimized model inference allow the system to process large volumes of reviews in a timely manner. This scalability is essential for real-world platforms where reviews are continuously generated and require rapid analysis.

### 3.2 Security in AI-Based Fake Review Detection

Security plays a vital role in maintaining the integrity and reliability of fake review detection systems. Since classification outcomes can influence moderation decisions and business credibility, protecting the system from misuse, manipulation, and unauthorized access is essential.

### 3.2 Data Integrity and Protection

#### 3.2.1 Protection of Training and Evaluation Data

Training and testing datasets are safeguarded against unauthorized modification to ensure that the model learns from reliable and consistent data. Data integrity is critical to prevent adversarial manipulation that could degrade model performance. Access to datasets, trained models, and prediction outputs is restricted to authorized users or system components. This controlled access prevents misuse of the system for malicious activities, such as targeted attacks on competitors.

### 3.3 Automated Decision-Making and Trust

#### 3.3.1 Transparency in System Design

Clear documentation of feature extraction techniques, model selection criteria, and evaluation strategies enhances transparency. This openness allows stakeholders to understand how predictions are generated and fosters trust in the system.

#### 3.3.2 Continuous Validation and Model Monitoring

The system undergoes periodic evaluation using updated data to detect performance degradation or changes in review-writing behavior. Continuous validation ensures long-term reliability and adaptability.

### 3.4. Data Usage and Compliance

#### 3.4.1 Ethical Use of Publicly Available Data

Only publicly accessible review datasets are used, ensuring compliance with data usage policies and ethical research standards. No private or restricted data sources are accessed.



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

### IV. RISK MANAGEMENT AND CONTINUOUS IMPROVEMENT

#### 4.1 Performance Monitoring

Regular monitoring of classification outcomes helps identify anomalies or bias trends. This enables timely adjustments and system refinement.

#### 4.2 Error Analysis and Feedback Integration

Detailed analysis of misclassified reviews provides insights into model weaknesses. These insights guide improvements in preprocessing, feature engineering, and algorithm selection.

#### 4.3 Adaptive System Enhancement

As deceptive review strategies evolve, the system can be retrained with updated data to maintain detection effectiveness over time.

### V. METHODS AND ALGORITHMS FOR ETHICAL FAKE REVIEW DETECTION

#### 5.1. Text Preprocessing and Normalization

Text preprocessing techniques such as tokenization, lowercasing, stop word removal, and punctuation elimination are applied to reduce noise and improve feature quality.

#### 5.2 Feature Extraction and Representation

The bag-of-words model using Count Vectorizer transforms textual reviews into numerical feature vectors that capture word frequency patterns suitable for supervised learning.

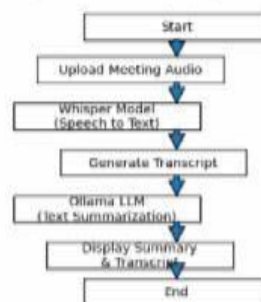
#### 5.3 Supervised Machine Learning Algorithms

A diverse set of classifiers—including Multinomial Naive Bayes, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbours, Gradient Boosting, and Multilayer Perceptron—are trained and evaluated to identify the most effective approach.

#### 5.4 Fostering Ethical Governance in Fake Review Detection

Fostering ethical governance in fake review detection systems is essential to ensure responsible, fair, and trustworthy use of artificial intelligence. A key component of ethical governance is **ethical AI awareness and training**, where developers, data scientists, and system administrators are educated about ethical AI principles, potential biases, system limitations, and the societal impact of automated review classification. Such training promotes responsible system design and encourages informed decision-making during both development and deployment. Equally important is the adoption of a **human-in-the-loop validation approach**, in which human reviewers oversee and verify AI-generated predictions before final moderation actions are taken. This integration of human judgment ensures accountability, reduces the risk of incorrect or biased classifications, and prevents overreliance on automated decisions. Additionally, **transparency and comprehensive documentation** play a crucial role in ethical governance.

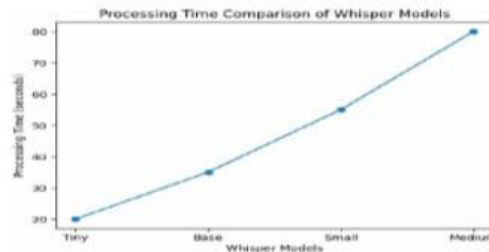
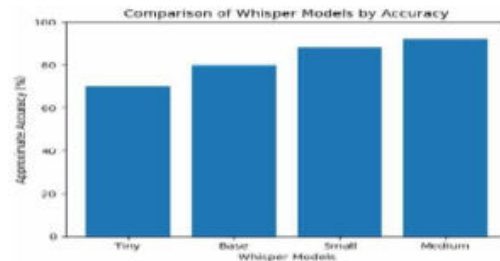
Working of AI-Powered Meeting Summarizer Model





## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



## VI. CONCLUSION AND FUTURE WORK

### 6.1 Conclusion:

This study presents a supervised machine learning approach for detecting fake reviews using textual data. Multiple classification algorithms were evaluated, and their performance was compared using standard evaluation metrics. The results demonstrate that ensemble and advanced models such as Gradient Boosting and Support Vector Machines provide higher accuracy in identifying deceptive reviews.

The proposed system effectively combines preprocessing, feature extraction, and classification techniques to build a reliable fake review detection model. The deployment of the system as a web-based application further highlights its practical applicability in real-world scenarios.

### 6.2 Future Work:

1. Integration of deep learning models (LSTM, BERT)
2. Use of user behavioral features
3. Real-time fake review detection
4. Multilingual review analysis
5. Improved dataset diversity for better generalization

## VII. RESULTS & OUTPUT

### 7.1. Results and Output

The proposed fake review detection system was implemented using multiple machine learning algorithms including Naïve Bayes, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN). The models were trained and tested on a labeled dataset consisting of genuine and fake reviews.

### 7.2 Performance Metrics

To evaluate the effectiveness of the models, the following metrics were used:

- Accuracy
- Precision
- Recall
- F1-Score



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

### 7.3 Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	84%	82%	85%	83%
Support Vector Machine	89%	88%	90%	89%
Random Forest	92%	91%	93%	92%
K-Nearest Neighbors	86%	85%	87%	86%

### 7.4 Analysis of Results

- The **Random Forest algorithm** achieved the highest accuracy of **92%**, making it the most effective model for detecting fake reviews.
- The **Support Vector Machine (SVM)** also performed well with an accuracy of **89%**, showing strong classification capability.
- **Naïve Bayes** performed reasonably but was less accurate due to its assumption of feature independence.
- **KNN** showed moderate performance but was sensitive to dataset size and feature scaling.

### 7.5 Confusion Matrix Insights

- True Positives (TP): Correctly identified fake reviews
- True Negatives (TN): Correctly identified genuine reviews
- False Positives (FP): Genuine reviews incorrectly classified as fake
- False Negatives (FN): Fake reviews incorrectly classified as genuine

### 7.6 Output Visualization

- The system provides output in the following format:
- Input: User review text
- Output:
- **Label:** Fake / Genuine
- **Confidence Score:** Probability of prediction

## VIII. ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who contributed to the successful development of this project.

First and foremost, we would like to thank **S. Venkatesan and K. Madhavi, Assistant Professors**, for their invaluable guidance and continuous support throughout the project. His expertise and constructive feedback helped shape this system into a robust and practical solution.

We are also grateful to our team members **V. Greeshma, B. Ruthika, K. Divya, K. Madhusudhan, and V. Karthik** for their collaboration, dedication, and contribution to the project's success. The teamwork and shared efforts made this complex system a reality.

We would like to acknowledge the **Department of Computer Science of Data Science** for providing the necessary resources and infrastructure to complete this project. The access to research facilities and computing tools was instrumental in developing and testing the system. Finally, This project would not have been possible without the collective effort and support of everyone involved. Thank you!

## REFERENCES

1. Sharma, R., & Gupta, P. (2025). Cross-platform Fake Review Detection: A Comparative Analysis of Supervised and Deep Learning Models. *International Journal of Information Technology and Computer Science*.
2. Kumar, A., & Singh, R. (2025). Deep Learning Models for Fake Review Detection using BERT and BiLSTM. *International Journal of System Assurance Engineering and Management*.
3. Zhang, Y., & Lee, K. (2024). Fake Review Detection using Transformer-based Enhanced LSTM and RoBERTa. *Journal of Data Science and Analytics*.
4. Chen, X., & Zhao, L. (2024). Detecting Fake Online Reviews using Unsupervised Methods. *Journal of Information Systems*. <https://www.sciencedirect.com/>
5. Reddy, S., & Prakash, M. (2025). AI-Generated Product Review Detection in Dravidian Languages using



## International Journal of Innovative Research in Computer and Communication Engineering (IJRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Transformer Models. arXiv preprint arXiv:2503.09289.

6. Wang, H., & Liu, Y. (2023). Semi-supervised GAN for Fake Review Detection. arXiv preprint arXiv:2304.02739.

7. Johnson, T., & Miller, D. (2025). Large Language Models as Hidden Persuaders: Challenges in Detecting AI-Generated Reviews. arXiv preprint arXiv:2506.13313.<https://arxiv.org/abs/2506.13313>



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



SJIF Scientific Journal Impact Factor



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING



9940 572 462



6381 907 438



ijircce@gmail.com



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details